Maintaining Consistency in extended Multiple text generation: An exploration of parametric Medical Personas

Aadit Deshpande
aaditd@andrew.cmu.edu

1 Introduction

Recent advances in large language models (LLMs) (Touvron et al. (2023), OpenAI (2023), Taori et al. (2023)) have produced models that can generate very fluent text and apparently seem to have a good understanding of both linguistic components (such as morphology and syntax) and semantic components. Naturally, there has been growing interest in the application of LLMs to writing and textual understanding tasks in the medical domain. Medical question answering has been heavily focused on, with techniques such as using an ensemble of language models (Yang et al., 2023) and retrieval training to enhance reasoning (Wang et al., 2024). Another area of active research is automatic medical SOAP note generation such as using LLMs that have been continually pre-trained on domain-specific text (Yuan et al., 2024) or using fine-tuned models with attribute-specific cross-attention parameters (Ramprasad et al., 2023).

SOAP notes (Subjective, Objective, Assessment, and Plan) are a widely adopted means of documentation for health providers in the United States. They are siginificant because they communicate crucial information about the health status of a patient and also act as a communication document among health professionals. Initially, this project focused on generating synthetic depersonalized SOAP notes that could then be used as data for downstream Information Retrieval (IR) tasks. Some of the requirements of such a synthetic SOAP note would include: depersonalization, sufficient detail, factual correctness, correlation of information among section, and that the four sections be related to each other. Additionally, the ability to parameteize the SOAP note generation would be highly beneficial for this use case. We tried two kinds of approaches. First, using a single model to generate each of the 4 sections at once. This proved difficult due to constraints on the context window length and the model's increased proclivity to hallucinate with increasing input length. Thus, we switched over to the other extreme i.e. generating each section independently. Though the hallucinations were reduced, the sections were often completely unrelated.

This, motivated us to switch gears to a different, less studied research question: 'how do we maintain consistency among such a "turn-wise" approach to multiple text generation?'. In the current study, the texts we chose to focus on are medical 'personas' and the blog posts they write about their daily lives. There are several reasons for choosing medical personas over SOAP notes, mainly because they are far more expressive, show variety in style, content, and format. This degree of flexibility allows for many avenues for us to discover common types of model errors. We consider consistency along three dimensions. First, Coherence: Is the generated text factually correct? Second, Plausibility: Is the current generated text consistent with the already established facts from previous generations? Finally, Time-scale: For how many turns can the model maintain coherence and plausibility? The third dimension of consistency is a major research focus of this project, wherein we ask the question: "What prompting strategy can we use to project scenarios into the future and to what time-scale can we reasonably do so?" Thus in the present study, our contributions are [1]:

- We propose a cascading summarizer-generator pipeline using the Llama2-7b-chat model that allows for the efficient generation of extended multiple texts (medical persona blog posts).
- We come up with an ontology of the most common types of errors and perform qualitative analysis of model performance.

[1] Code and generations are available at https://github.com/Aadit3003/llm-medical-personas

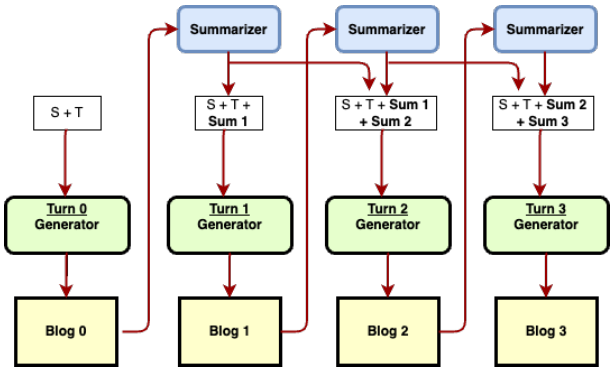

Figure 1: The cascading summarizer-generator architecture used to generate the blogs for the personas. S and T stand for the System and Task prompts fed to the generator model respectively, and Sum_n stands for the n^{th} blog summary.

- We propose five new metrics (Original Correctness/Recall, Novel Correctness/Recall, New Facts per turn) to possibly evaluate the consistency of long-range texts generated over multiple turns.

2 Method

2.1 Problem Setup

We operationalize medical personas using "Universal Fact Lists" - Lists of attributes that can take different values allowing customization of medical and personal profiles. We consider attributes such as the medical condition, its severity, the patient's age, gender, profession, location, what medications they take, any activity restrictions imposed on them, allergens, and any food or substance restrictions. Table 4 in Appendix A provides examples of such Universal Fact lists.

2.2 Architecture

We propose a cascaded summarizer and generator architecture as shown in Fig 1 for the turn-wise generation of blog posts with respect to a consistent medical persona. The generator model generates a blog post from the prompt (described in the following section). Since the generator has a limited context window, we employ a summarizer model to shorten the blog post and append it to the prompt for the next blog generation. Our qualitative analysis revealed that this greatly improves the consistency among the blog posts, particularly the plausibility. After much experimentation, we found that the best combination of models was as follows:

- Generator: Llama2-7b-chat-hf
- Summarizer: Both these options work well and produce different kinds of summaries.
  - Abstractive Summarizer: Llama2-7b-chat-hf (One-shot with task description)
  - Extractive Summarizer: Maximal Marginal Relevance (MMR)

Surprisingly the Maximal Marginal Relevance summarizer (a technique used in Information retrieval to diversify search engine rankings Carbonell and Goldstein, 1998) works a highly efficient and effective extractive summarizer, by selecting sentences in a way that minimizes information loss as well as redundance. We employ a past-look-over of $p = 2$, i.e., the generator model is allowed to only look at the two most recent blog post summaries when it generates a new blog post.

2.3 Prompting

Table 1 shows the prompting strategy we use for the generator model. First, we use the System prompt (S) to induce the medical persona in the model, followed by the Universal Fact List. Next, in the Summary Prompt (Sum) we concatenate the summaries of the preceding two blog posts, and finally we append the Task prompt (T) to get the model to generate the blog post. The System prompt and Task prompt remain the same and are included in every turn, whereas the Summary prompt changes dynamically. Thus, at turn n (≥ 2), the prompt is:-

$$blog_{n-1} = Generator(prompt_{n-1})$$

$$Sum_{n-1} = Summarizer(blog_{n-1})$$

$$prompt_n = S + Sum_{n-1} + Sum_{n-2} + T$$

2.4 Experimental Design

We conducted a qualitative and quantitative study on maintaining consistency among eight distinct personas for a chronic condition, namely Asthma. This is because it's widely known and is common enough on the web to have occurred in the pretraining data for Llama2-7b enough times for it to have a sophisticated understanding of the condition. The personas were crafted to have minimal overlap and show diversity in age, profession, gender to induce stylistic differences in the model's generations. Studying multiple personas has several advantages- First, it helps average out any noise

Prompt Type	Prompt
System Prompt (S)	You are a 30 year-old woman with asthma. You work as an actor, writing your monthly blog based on the interesting events in your life. Write about how the following facts change over time: <universal_fact_list>
Summary Prompt (Sum)	This is a summary of all your previous blog posts: <previous_blog_summary>
Task Prompt (T)	Now write a blog post for the month of April. Blog Post:

Table 1: The prompt structure used in the architecture with the Llama2 generator models.

in the metrics that might occur due to an exceptional or quirky persona. Second, it allows for an investigation of consistency not only in content, but also in the style of the generated text.The details of these personas are listed in Table 4 in Appendix A. We use the MMR summarizer, and a past-look-over of $p = 2$. Additionally, in order to study whether the models remember long temporal dependencies, we set the time-period granularity to 1 month i.e. the model generates monthly blogs for the eight distinct personas.

3 What did Not work

Before proceeding to the experimental analyses, we very briefly discuss certain approaches that we tried, but did not yield any significant improvements to generation or summarization quality, for the benefit of readers and future researchers interested in improving upon extended text generation.

3.1 Summarization models

Initially, we used the FlanT5-xl-3B and BART-Large-CNN models for summarization, however they both yielded poor summaries which in turn led to disastrous effects on all future generations. With FlanT5 summaries, the main issue was information loss, as the model would often produce one sentence summaries that would not capture the detail necessary to prompt the following blog post generation. With BART-Large, the model instead defaulted to returning the first three sentences of the blog as the summary. While this can be a useful heuristic for news corpora, it is virtually useless for the blog post case, as detail is not usually concentrated towards the start of the article.

3.2 Fact checking models

We also experimented with NLI models such as BART-large-mnli and mDeBERTa-v3-base to extract facts from the current generated text and evaluate their plausibility with respect to the Universal Fact List (framed as an entailment problem). However, this approach did not seem to work, as the NLI models were highly biased towards the 'Neutral' option. Removing the neutral option and selecting the 'Entail/Contradict' option was no better than random guessing. The models also struggled a lot with numericals such as not being able to differentiate between examples like comparing a diagnosis of 'six years' and 'six months'.

3.3 Specialized Medical models

Before switching to the base Llama2-7b, we initially started by using Meditron-7b (Chen et al., 2023) which is Llama2 that has been continually pre-trained on PubMed articles and medical guidelines. While the Meditron models show impressive performance on medical QA benchmarks, it is not well-suited to the current text generation task. The biggest issue with the Meditron model was that its generations almost followed the format of a research paper rather than a blog, making it very difficult to prompt, compared to the more flexible base Llama2.

4 Qualitative Error Analysis

In this section, the larger question we want to address is whether the source of errors the model makes is due to it 'forgetting' with time, or the fact that the model cannot recognize contradictions. Our results show that both are viable explanations for the errors the model makes. To this end, we come up with an ontology of errors to categorize the most common mistakes. Some examples are shown in Table 2. The four broad categories of errors (in descending order of seriousness) are:-

- Type-A: The model gets a medical fact factually incorrect. For example talking about genes that affect prostate cancer for a patient

Persona	Blog Post	Error	Notes
30 y/o woman, Actor, Severe Asthma, CA	• Blog 5:-First off May marked 5 years since diagnosis! This was particularly significant because since then I haven't required emergency hospitalization due to severe exacerbations • Blog 6:- Today marks seven years since diagnosis, and boy do things feel different! Let me tell you, being nearly three decades old with adult asthma isn't fun • Blog 7:- I find myself reflecting back on time this whole thing started. It was eight long years ago now since first detected & properly diagnosed with Adult Onset Asthma...and let me tell you IT HAS BEEN a ROLLER COASTER!	Type-B	Numerical Inconsistency
40 y/o woman, teacher, Severe Asthma, NY	• Blog 1-7: *<no mentions of myasthenia gravis>* • Blog 8: Title: Breathless at Midyear Body Text:The dog days of summer have arrived leaving many of us reaching for our Ventolin or other quick relievers As I sit here typing away my computer screen fan blowing directly onto sweaty face I'm thinking back over last six months since my January post when I discussed the various ways myasthenia gravis (AG) was affecting me.	Type-C	Mentioned once and never again
20 y/o man, college, Severe Asthma, CA	Blog 11: "Winter Blues Got You Down? Here's Tech Solutions to Help Lift Your Spirits"In this piece explore latest technologies designed to enhance mental well being winter months. Some key areas focus on include mood tracking wearables meditation apps, therapy apps,music streaming services.	Type-D	Unexpected Domain Shift

Table 2: Qualitative examples of the most common types of errors made by the model while generating the blog posts.

Metric	**Turn**											
	1	2	3	4	5	6	7	8	9	10	11	12
OC	0.66	0.53	0.48	0.54	0.48	0.43	0.44	0.52	0.41	0.42	0.48	0.42
OR	0.45	0.36	0.31	0.29	0.34	0.27	0.24	0.33	0.27	0.29	0.24	0.24
NF	1.13	1.25	0.88	1.38	1.13	1.75	1.25	1.38	1.88	1.38	1.25	0.88
NC	0.50	1.00	0.63	0.81	1.00	0.58	0.56	0.83	0.84	0.63	0.46	0.44
NR	0.50	0.78	0.25	0.39	0.30	0.35	0.14	0.16	0.22	0.16	0.08	0.07

Table 3: The newly proposed metrics averaged across eight distinct personas (to reduce noise in the data). OC = Original Correctness, OR = Original Recall, NF = New Facts per turn, NC = Novel Correctness, NR = Novel Retrieval.

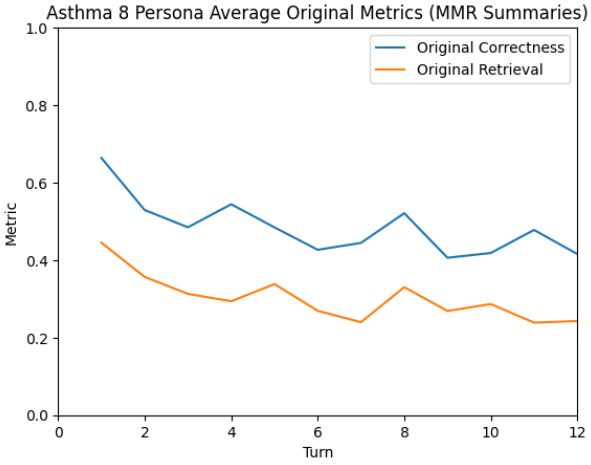

Figure 2: The Original Correctness and Retrieval metric curves (averaged across 8 personas) plotted across 12 turns.

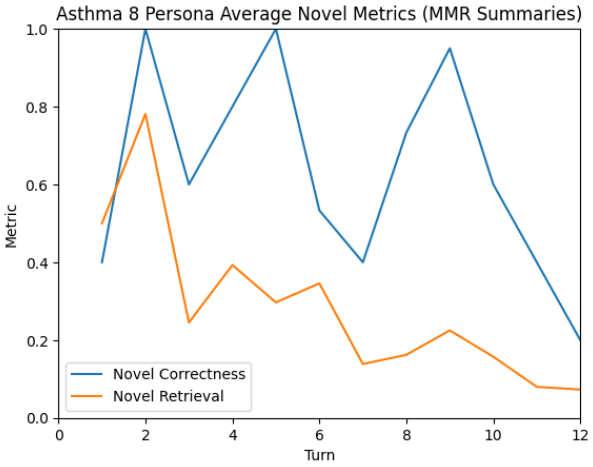

Figure 3: The Novel Correctness and Retrieval metric curves (averaged across 8 personas) plotted across 12 turns.

who was assigned female at birth. These errors are difficult to identify without the help of a human annotator with medical expertise.

- Type-B: Inconsistency with established Constraint. For example the model makes a mistake by contradicting one of the location/allergen/medication constraints in the universal fact list. The most frequent occurrence of this type of error seems to be with numericals, as Llama-2 struggles with numerical understanding especially with respect to time. For e.g. Table 2 shows an example whether successive blog posts are unable to identify that they're making errors by constantly rewriting the years since diagnoses for a certain medical persona.

- Type-C: A new plausible fact is mentioned once but never seen again in future generations. This type of error is less of a threat to plausibility of blog posts, but is also very common. The most frequent manifestation of this in our experiments happened when the personas would change medication doses, or be diagnosed with a disease, only for the new change to be forgotten almost immediately in the subsequent blog posts.

- Type-D: Finally, this type of error involves an unexpected domain shift which brings in irrelevant facts to the medical profile. This tends to occur later in the generation process, at about turns 10 or 11, and usually the blog posts about asthma suddenly transition into a generic text about 'mental well-being' with no reference to the established constraints or medical condition.

Overall, the Type-B errors (inconsistency with established constraint) seem to be the most common, and most threatening to the plausibility dimension of consistency. An additional issue is whether the Generator model can successfully preserve the style of writing across all turns. From our results and human evaluation, this highly depends on the details of the persona as well as the summary quality. Younger personas are easier for the model to maintain stylistic consistency for, whereas older personas tend to suddenly change writing styles every 1-2 blog posts (such as the overuse of emojis, uncharacteristic use of slang or abbreviations associated with younger demographics). However, this is out of scope of the current study and needs deeper analysis.

5 Proposed Metrics and Quantitative Analysis

In order to answer the question about how many turns we can sustain consistency, we propose five new consistency metrics operationalized as follows:-

- Original Correctness: The proportion (weighted) of original facts (from the universal fact list), that the current blog post can be used to answer correctly with a Question-Answering model.

- Original Retrieval: The proportion (wieghted) of original facts retrieved in the blog post.
- New Facts per turn: The number of new facts introduced in the blog post at each turn.
- Novel Correctness: The proportion of new facts at each turn that is plausible (consistent with the original facts in the universal facts list). This metric rewards novelty and penalizes contradictions. In other words, it incentivizes the model to generate many new plausible facts.
- Novel Retrieval: The number of novel facts retrieved at each turn in the blog. This metric rewards novelty, but harshly penalizes forgetting.

In the above definitions, by 'weighted', we mean that each attribute in the universal fact list is assigned a relative weight (Table 4) such as allergens and medications having higher weight than location. In the current pilot study, only Original Correctness is automated (measured by the QA model), and the other four metrics are calculated manually (by analyzing 8 personas x 12 turns) due to time constraints. Finding ways to automate the calculation of metrics is a promising direction of future research. We average the metrics across the eight personas, as shown in Table 3 to reduce the noise and variability in the metrics.

Fig 2 shows the Original Correctness (OC) and Retrieval (OR) metrics plotted across the 12 turns. These metrics are analogous to Precision and Recall. OC tends to have hgiher values than OR and both tend to have only slightly decreasing trends. This is a good sign, as this means that model performance with respect to maintaining consistency with original facts does not degrade significantly with increasing number of turns.

Fig 3 in contrast shows the Novel Correctness (NC) and Retrieval (NR) metric plots. These curves are significantly noisier, despite being averaged across eight different trials (personas). NR seems to be less noisy and decreases more sharply than NC. This is because the list of novel facts is dynamic and keeps growing, unlike the static universal facts list, making recall a more difficult task with increasing turns. The high amount of noise in NC can be explained by the fact that in most turns, the blog introduces only 1-2 new facts (Table **??** and Table 9), meaning it is assigned a 0/1 score causing a lot of fluctuations in the plot.

6 Conclusion

In the present study, we examine the question - how can we prompt LLMs to maintain consistency (plausibility and coherence) over extended multiple text generation and we particularly focus on medical personas. We find that a cascaded generator + summarizer architecture is highly effective, including using MMR (Maximal Marginal Relevance) for summarization and Llama2-7b for generation. We conduct an experiment by creating eight parametrizable medical personas for asthma. We conduct a qualitiative analysis of the model's errors and propose an ontology of common error types. Summary quality seems to be the biggest determiner for the goodness of all future generations. We also investigate the question about how far into the future LLMs can maintain consistency (timescale). To this end, we propose 5 new metrics - Original Correctness/Retrieval, New Facts per turn, and Novel Correctness/Retrieval. We observe that the Original Correctness and Retrieval stay stable with only a slightly decreasing trend, indicating that LLMs are capable of maintaining consistency over text generation and that this is a promising area for research.

Future studies could focus on issues such as automating the calculation of the metrics, studying more medical conditions and comparing acute and chronic conditions, even more attributes to customize the parametrizable personas, and examining stylistic consistency in addition to content consistency.

References

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

R OpenAI. 2023. Gpt-4 technical report. *ArXiv*, 2303.

Sanjana Ramprasad, Elisa Ferracane, and Sai P Selvaraj. 2023. Generating more faithful and consistent soap notes using attribute-specific parameters. In *Machine Learning for Healthcare Conference*, pages 631–649. PMLR.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Junda Wang, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024. Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability. *arXiv preprint arXiv:2402.17887*.

Han Yang, Mingchen Li, Huixue Zhou, Yongkang Xiao, Qian Fang, and Rui Zhang. 2023. One llm is not enough: Harnessing the power of ensemble learning for medical question answering. *medRxiv*.

Dong Yuan, Eti Rastogi, Gautam Naik, Jai Chintagunta, Sree Prasanna Rajagopal, Fen Zhao, Sagar Goyal, and Jeff Ward. 2024. A continued pretrained llm approach for automatic medical note generation. *arXiv preprint arXiv:2403.09057*.

A Appendix

This section contains the following interesting helpful supplementary information that was too large to include in the main text, and may help interested readers reproduce our experiments:

- Table 4 and Table 5 contain details about the Universal Fact Lists of eight distinct medical personas for Asthma
- Table 6 and Table 7 contain the Original Correctness/Recall metrics for each individual persona as well as their averaged results across the 12 turns.
- Table 8, Table 9, and Table 10 contain the New Facts per turn, Novel Correctness, and Novel Recall metrics respectively for each individual persona as well as their averaged results across the 12 turns.

S.No.	Case	Age	Gen	Prof (2)	Location (3)	Med (4)	Activity Restrictions (3)	Allergens (5)	Substance Restrictions (5)
P1	Severe 1	30	W	Actor	Pasadena, CA	S1	<45 mins, No cold-weather	Pets	Alcohol
P2	Severe 1	20	W	Athlete (Swimmer)	Los Angeles, CA	S1	<1 hr, Outdoor air pollution	Pollen	Alcohol + Fatty processed foods
P3	Severe 2	40	W	Teacher	New York, NY	S2	Outdoor air pollution	Fragrances	Fatty processed foods
P4	Mild	70	W	Retired	Miami, FL	M	Dry weather, high humidity	Dust-mites	nonsteroidal anti-inflammatory drugs
P5	Severe 2	20	W	College Student	Wisconsin, Madison	S2	<45 mins, No Cold-weather	Fragrances	Vaping, Smoking
P6	Severe 1	20	M	College Student	Los Angeles, CA	S1	Pollution	Pets	Smoking + Alcohol
P7	Severe 2	30	M	Actor	New York, NY	S2	Pollution	Pollen	Alcohol + nonsteroidal anti-inflammatory drugs (NSAIDs)
P8	Mild	40	M	Construction Worker	Pittsbrugh, PA	M	<45 mins, No cold weather exposure	Dust-mites	Fatty processed foods, smoking

Table 4: Specification of the eight distinct asthma personas considered in the study. Customizable attributes include severity of case, gender, profession, location, medication, activity restrictions, allergens, and food/substance restrictions. Medication sets (M, S1, S2) are specified in Table 5. The numbers in parentheses indicate the relative weight assigned to each attribute while calculating the Original Correctness metric.

Asthma Case	Medication Set	Detail
Mild	M	Pulmicort Flexhaler and Albuterol
Severe 1 (somewhat severe)	S1	Pulmicort Flexhaler and Perforomist inhalation solution (LABA)
Severe 2 (extremely severe)	S2	Pulmicort Flexhaler and Vilanterol Inhalation Powder

Table 5: Details about medication sets depending on severity of asthma.

Persona	Turn											
	1	2	3	4	5	6	7	8	9	10	11	12
P1	0.94	0.56	0.56	0.34	0.31	0.66	0.56	0.56	0.56	0.41	0.53	0.41
P2	0.70	0.68	0.22	0.54	0.62	0.62	0.46	0.30	0.22	0.22	0.30	0.30
P3	0.66	0.66	0.34	0.52	0.66	0.28	0.38	0.52	0.52	0.66	0.72	0.66
P4	0.41	0.52	0.24	0.34	0.10	0.28	0.28	0.59	0.10	0.59	0.34	0.52
P5	0.69	0.47	0.53	0.59	0.41	0.31	0.41	0.41	0.38	0.41	0.41	0.34
P6	0.62	0.47	0.59	0.59	0.35	0.35	0.50	0.79	0.21	0.09	0.50	0.38
P7	0.65	0.29	0.44	0.41	0.59	0.41	0.44	0.62	0.59	0.44	0.62	0.44
P8	0.70	0.62	0.49	0.78	0.62	0.49	0.62	0.62	0.62	0.65	0.49	0.38
Average	0.66	0.53	0.48	0.54	0.48	0.43	0.44	0.52	0.41	0.42	0.48	0.42

Table 6: The **Original Correctness** metric for the eight personas (P1-P8) across 12 turns and the averaged values.

Persona	Turn											
	1	2	3	4	5	6	7	8	9	10	11	12
P1	0.76	0.66	0.34	0.00	0.55	0.45	0.48	0.34	0.38	0.34	0.24	0.24
P2	0.26	0.26	0.47	0.26	0.41	0.38	0.26	0.00	0.26	0.21	0.06	0.00
P3	0.48	0.34	0.38	0.24	0.38	0.38	0.17	0.72	0.45	0.48	0.24	0.45
P4	0.66	0.31	0.24	0.41	0.00	0.00	0.00	0.41	0.00	0.00	0.38	0.00
P5	0.34	0.31	0.24	0.45	0.59	0.00	0.31	0.00	0.31	0.38	0.24	0.52
P6	0.32	0.41	0.32	0.47	0.38	0.32	0.29	0.38	0.21	0.21	0.21	0.21
P7	0.35	0.21	0.21	0.32	0.21	0.32	0.21	0.59	0.35	0.32	0.35	0.21
P8	0.38	0.35	0.30	0.19	0.19	0.30	0.19	0.19	0.19	0.35	0.19	0.32
Average	0.45	0.36	0.31	0.29	0.34	0.27	0.24	0.33	0.27	0.29	0.24	0.24

Table 7: The **Original Retrieval** metric for the eight personas (P1-P8) across 12 turns and the averaged values.

Persona	Turn											
	1	2	3	4	5	6	7	8	9	10	11	12
P1	2	2	1	0	1	3	1	1	1	1	0	0
P2	1	2	1	2	1	4	2	1	2	0	1	0
P3	0	1	0	2	1	2	0	1	3	2	0	2
P4	4	1	3	1	1	0	1	3	4	0	3	1
P5	0	1	0	2	2	0	1	2	1	2	0	0
P6	1	1	1	1	1	3	1	1	2	2	3	1
P7	1	1	0	1	1	2	2	1	2	2	3	2
P8	0	1	1	2	1	0	2	1	0	2	0	1
Average	1.13	1.25	0.88	1.38	1.13	1.75	1.25	1.38	1.88	1.38	1.25	0.88

Table 8: The **New Facts per turn** metric for the eight personas (P1-P8) across 12 turns and the averaged values.

Persona	**Turn**											
	1	2	3	4	5	6	7	8	9	10	11	12
P1	1.00	1.00	1.00	0.00	1.00	0.67	0.00	0.00	1.00	1.00	0.00	0.00
P2	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	0.00
P3	0.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	0.00	1.00
P4	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.67	0.75	0.00	1.00	0.00
P5	0.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	0.00	0.00
P6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.50	1.00	1.00
P7	1.00	1.00	0.00	1.00	1.00	1.00	0.50	1.00	1.00	0.50	0.67	0.50
P8	0.00	1.00	1.00	0.50	1.00	0.00	1.00	1.00	0.00	1.00	0.00	1.00
Average	0.50	1.00	0.63	0.81	1.00	0.58	0.56	0.83	0.84	0.63	0.46	0.44

Table 9: The **Novel Correctness** metric for the eight personas (P1-P8) across 12 turns and the averaged values.

Persona	**Turn**											
	1	2	3	4	5	6	7	8	9	10	11	12
P1	1.00	0.50	0.20	0.00	0.17	0.25	0.00	0.00	0.11	0.10	0.00	0.00
P2	0.00	1.00	0.33	0.50	0.25	0.67	0.13	0.11	0.20	0.00	0.09	0.00
P3	0.00	1.00	0.00	0.67	0.25	0.60	0.00	0.17	0.43	0.22	0.00	0.22
P4	1.00	0.25	0.43	0.14	0.13	0.00	0.00	0.20	0.25	0.00	0.21	0.00
P5	0.00	1.00	0.00	0.67	0.50	0.00	0.20	0.29	0.14	0.22	0.00	0.00
P6	1.00	1.00	0.50	0.33	0.33	0.75	0.25	0.20	0.33	0.29	0.33	0.11
P7	1.00	0.50	0.00	0.50	0.50	0.50	0.20	0.17	0.33	0.14	0.00	0.13
P8	0.00	1.00	0.50	0.33	0.25	0.00	0.33	0.17	0.00	0.29	0.00	0.13
Average	0.50	0.78	0.25	0.39	0.30	0.35	0.14	0.16	0.22	0.16	0.08	0.07

Table 10: The **Novel Retrieval** metric for the eight personas (P1-P8) across 12 turns and the averaged values.